Detection of polyphonic music note onsets by application of the Bayesian Theory of Surprise

Piotr Holonowicz, Perfecto Herrera

Music Technology Group, Departament of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain

piotr.holonowicz@upf.edu, perfecto.herrera@upf.edu

Abstract

In this paper we present an onset detection algorithm that consists of two parts, the detection of transient peaks in an audio spectrum and the classification of the peaks, adapting a model derived from the Bayesian Theory of Surprise. The model is an unsupervised, robust adaptation of conjugate priors, providing the distributions of beliefs about the number of the transient peaks, in a time space as well as in a frequency space. The novelty points marked by the model are then classified according to their relevance in order to filter out non-onset events, caused for example by a background noise. It has been evaluated using a collection of over 170 music excerpts. Our experiments show that the new model can provide an overall performance close to the current state of the art solutions. We discuss the advantages of the presented approach and the ways to overcome its shortcomings and the possible directions of future research.

Index Terms: onset detection, Bayes, modeling, surprise, novelty

1. Introduction

The detection of the beginnings of the notes in digital music streams is one of the fundamental problems in the Sound and Music Computing field, as it provides a partition of the audio stream into the smallest separate segments that have musical meaning. A note onset may be characterized as a change of the sound that is relevant in the context of our previous listening experience. For polyphonic music, an onset may be defined as the start of the attack phase of one or more simultaneous sounds if they are perceptually perceived as one note, or a fluent change of at least one of the perceived pitches. In the last case the onset can be definied as a period of time instead of a single point in the time axis.

The detection of the onset events is generally realized using algorithms working either in the time domain or in the spectral domain. Time-domain algorithms are mainly looking for sharp slopes of the instantaneous energy and they are robust only for percussive sounds [1]. However, they are still used for speech recognition as they provide the time resolution impossible to achieve by algorithms working in the spectral domain [2]. For analysing music the Short Time Fourier Transform based solutions are widely used, providing a possibility of modeling perceptual features like frequency masking. The design principle of these solutions is a construction of an onset detection function which is then thresholded for obtaining the onset times. The comprehensive description of the spectral based onset detection algorithms can be found in [1].

A new family of methods combining spectral-based processing with machine learning techniques has been intensively investigated recently, probably motivated by the success of the Hidden Markov Models(HMM) in speech recognition [3]. An interesting solution based on HMM's and Independent Component Analysis has been presented by Abdallah [4]. However, a better evaluated and more successful approach seems to be that by Lacoste & Eck, reaching average F measure = 0.79 at the MIREX 2007 [5].

An example of combining an energy-based method working in the spectral domain and a probabilistic model is the algorithm of onset detection proposed by Röbel [6]. We decided to make it a base of our approach because of its very good performance and the possibility of extension of the probabilistic part which, in our opinion, deserved a more comprehensive (or complex) approach.

The solution proposed in this paper consists of two parts, the detector and the classifier of the transient peaks. The block diagram of the system is presented at Figure 1. The transient peaks are detected with the same method as in Röbel system[6], by thresholding their Centers of Gravity. Further, the transient peaks are assigned into frequency bins and the amount of surprise is computed as a distance between the distribution of the number of the transient peaks within a frequency bin and the prediction done by a Bayesian model [8]. As the increased surprise does not always point to an onset, a two stage relevance filter has been introduced. The outcome of the system are the times of the onset events present in the input audio stream.

The paper is structured as follows: the original detector is described in Section 2, whereas Section 3 contains the description of our solution. The performance of our system has been evaluated on a data set that provides a wide variety of real music recordings, coming from different genres (see Section 4). The discussion about the results and the direction of the future research are described in Section 5.

2. Original model by Röbel

2.1. Centers of gravity

Following [6], the standard Short Time Fourier Transform (STFT) returns the magnitude spectrum A and the phase spectrum ϕ of a windowed signal. If the signal $s(t)$ is windowed with the analysis window $h(t, t_m)$ centered at the time position t_m, the spectrum is

$$S_h(\omega, t_m) = A(\omega, t_m)e^{j\phi(\omega, t_m)} \quad (1)$$

where ω is the frequency in radians.

The Center of Gravity (*COG*) of a peak can be then defined as the position (in time) of the centroid of the temporal energy

distribution in a given time-frequency region:

$$t_{cg} = \frac{\int_{\omega_l}^{\omega_h} -\frac{\partial\phi(\omega,t_m)}{\partial\omega} A(\omega,t_m)^2 d\omega}{\int_{\omega_l}^{\omega_h} A(\omega,t_m)^2 d\omega} \quad (2)$$

Please note that the above estimation of the *COG* operates local in frequency, which means that the integrals are limited to the frequencies located around the peaks (ω_l and ω_h are positions of the local minima of the amplitude, see [6] for the derivation of the formula and the details).
If the attack of an instrument is modeled as a single transient sinusoid that has the amplitude envelope shaped as a linear ramp with saturation, and we move a sliding window from the left over it, the *COG* moves "up" during the attack and "down" during the saturation (see [6], Figure 1). The movement is dependent on the slope of the attack. Thus the attack can be detected by thresholding of the motion curve of the *COG*. The threshold has been set empirically to $C_e = 0.1479$, so the transient is close to the signal center if the *COG* is close to C_e. The maximum *COG* still exceeds C_e even for a stationary sinusoid with amplitude only 20 dB above the background noise [6].

2.2. Binomial classification of the transient peaks

Spectral peaks related to the signal attacks have the *COG* far off the center of the window, however, a noise may cause non-transient signals to share this property, too. Fortunately, the peaks related to the attacks are mutually synchronized and the synchronization of a sufficient number of them is the criteria that allows to distinguish those from the peaks related to noise [6].
Röbel has proposed a binomial model describing the probability of a spectral peak p to have $COG > C_s = KC_e$ with $K \geq 1$ [6]. K is a parameter that controls the sensivity of the model thus it has a major influence on the robustness of the detection. The model requires the number of independent events N as a parameter it is important if N is not bound to the confidence of the decision, as the transients with single, wideband peak would be biased by the low number of observed peaks. Therefore it cannot be a number of peaks belonging to a frequency band, although this choice would be natural. Instead, an average number of peaks that might be contained in a band given the analysis window, is proposed as N value. As mentioned in 3.1, the model operates with the data coming from a spectrum divided into overlapping frequency bands, spaced equally, thus the desired transient probability must be consistent with the number of transient hits n in a frequency band within the range of G times the standard deviation of the mean value pN. Therefore it is required that

$$n = pN \pm G\sigma = pN \pm G\sqrt{p(1-p)N} \quad (3)$$

The model computes the transient probability p_c for the current F_c frames and compares it versus the transient probability p_h in the last F_h frames. Solving then equation (3) for p we obtain

$$p_c = \frac{G^2 N_c + 2n_c N_c - G\sqrt{N_c(G^2 N_c + 4n_c N_c - 4n_c^2)}}{2N_c(G^2 + N_c)} \quad (4)$$

N_c and n_c are the number of independent events and observed transient peaks for the current F_c frames. Probability p_h we obtain by replacing p_c, N_c and n_c in the formula 4 by the number of independent events and observed transient peaks in the frame history (p_h, N_h and n_h, respectively). An attack transient is detected if $p_c > p_h$ in any frequency band [6].
Although the model performed very well during the MIREX evaluations, it suffers from detection of too many false positives [6]. We tested the behavior of the *COG's* on test signals, consisting of pure sinusoids, a noise of various shapes and isolated sounds of various instruments. The model turns out to be sensitive to the low frequency modulations of sinusoidal signals (eg. vibratos). Another problem is that during the decay phase the SNR of a music signal becames very low. The symptom is the large number of random transient peaks spread across the frequency bands, that are hard to filter out with the model described above.

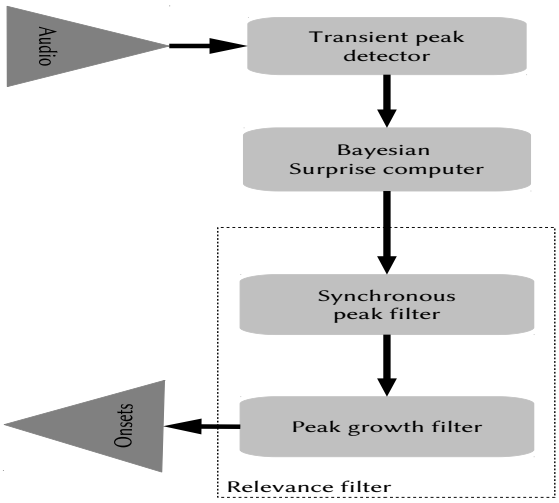

Figure 1: *Block diagram of the surprise-based onset detector.*

3. Onset detector based on Bayesian Surprise

3.1. Partition of the spectrum

The system presented in [6] divides the spectrum into overlapping, equally spaced frequency bands, so each band forms a "bin" which contains a number of transient peaks. In our opinion this scheme does not take into account the fact that the distribution of the frequencies in music signals is generally not uniform.
Being inspired by the ideas coming from the algorithms used for multi pitch estimation [7], we decided to propose an alternative partition, where the audible range 20 Hz - 20 kHz is partitioned into unequally spaced bands. Each one has the center in the frequency of a note belonging to the tempered scale. The borders of the bands f_l, f_u (lower, upper) are

$$f_l(i) = \frac{f_c(i) - f_c(i-1)}{2}, i \in [1, B-1] \quad (5a)$$

$$f_u(i) = \frac{f_c(i+1) - f_c(i)}{2}, i \in [1, B-1] \quad (5b)$$

where f_c is the center frequency of the band, i is the index of the band and B is the overall number of the bands. The center frequency can be expressed as

$$f_c(i) = f_{ref} \sqrt[12]{i - i_{ref}} \quad (6)$$

The f_{ref} is the A4 (the reference pitch, 440 Hz), i_{ref} is the corresponding, 49th key from the left end of a piano.
The numbers of peaks N_p and the number of transient peaks N_t are computed for each band separately. However the number of the bands is relatively high ($B \approx 170$). To reduce the dimensionality of the output, N_p and N_t are summed up for the frequencies corresponding to the h harmonics respective to each piano key. The number of harmonics is a variable dependent on the average RMS energy E_{dB} of the spectrum in each frame:

$$h = \lfloor \frac{E_{dB} + 90}{20} \rfloor \tag{7}$$

In consequence, the weakest harmonics are then not taken into account as being potentially noisy. The output of the preprocessing stage described above is then 88 pairs of values (N_t, N_p) for each frame.

3.2. The Bayesian Theory of Surprise

Onsets of notes in a music stream may be visualized as prominent changes in the FFT spectrum. When compared to the number of the STFT frames the typical audio signal is divided into, frames where onset events are present are relatively rare. In between, through a majority of time, we would expect that the spectrum would remain much less variable in time and the distribution of the number of transient peaks remains moreless constant within a frequency band [6]. The onsets may be then treated as surprising (novel) events. Itti & Baldi [8] propose to measure the degree of surprise by computing a Kullback-Leibler divergence [9] between the distribution representing prior beliefs of a family of models and the distribution representing beliefs of the models after exposure to the data

$$S(D, \mathcal{M}) = \int_{\mathcal{M}} P(M|D) \log \frac{P(M|D)}{P(M)} dM \tag{8}$$

where $\mathcal{M}$ is a family of models, M is a model representing a belief and D is a random variable representing the input data. Itti & Baldi [8] have compared the measure expressed in equation (8) with other metrics of novelty and they claim the proposed measure is the most consistent with observations made on humans. Although their experiment proved a successful application of the measure for computer vision, the surprise is formulated generally enough to be used in different domains as well, requiring only that the prior $P(M)$ and the posterior $P(M|D)$ distributions are available.

3.3. Beta-Binomial Model

The model described in Section 2.2 compares the current state with the state from a relatively short past (a few hundreds of miliseconds) [6]. Moreover, the binomial distribution essentialy models a sequence of independent trials. Thus, no long-time dependencies can be represented by the system. In hope to improve the accuracy of the detection, we propose a Bayesian solution which holds a memory of all the past states.
The number of transient peaks within each band is traced by a model consisting of a binomial node that represents our observations and a hidden Beta node representing beliefs about the transient peaks in each frame. The Beta distribution has been chosen because it is a conjugate prior to the binomial distribution. That eliminates the necessity of implementing complex methods to perform the inference, eg. Markov Chain Monte Carlo, as the conjugate posterior has the same form as the prior.

If P_M represents the model beliefs about the number of transient peaks in a frame and P_D represents the distribution of transient peaks in the frame k

$$M \sim Beta(\alpha_k, \beta_k), \tag{9a}$$
$$D \sim Binom(N_t, N_p - N_t) \tag{9b}$$

The density of the prior P_M is dependent on the two parameters

$$P_M(\alpha_k, \beta_k) = C_k x^{\alpha_k - 1}(1-x)^{\beta_k - 1} \tag{10}$$

where

$$\alpha_k \geq 0, \quad \beta_k \geq 0, \quad x \in [0, 1],$$
$$\alpha_k + \beta_k > 0, \quad C_k = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}.$$

Γ is the Gamma function. With the N_t transient peaks and $N_p - N_t$ non-transient peaks in a frame we may use the update rule for a conjugate Beta-Binomial pair

$$\alpha_{k+1} = \alpha_k + N_t, \quad \beta_{k+1} = \beta_k + (N_p - N_t). \tag{11}$$

The posterior probability distribution of the transient peaks is then $P_M(\alpha_{k+1}, \beta_{k+1})$. In case of the conjugates, the surprise function given by Eq. (8) can be derived exactly from Eq. (10) and Eq. (11) [12].

$$\begin{aligned} S(D, M) = \log \frac{C_k}{C_{k+1}} + N_t[\Psi(\alpha_k + \beta_k) - \Psi(\alpha_k)] \\ + (N_p - N_t)[\Psi(\alpha_k + \beta_k) - \Psi(\beta_k)] \end{aligned} \tag{12}$$

where Ψ is the derivative of the logarithm of the Gamma function. So in each frame for each band, the models are updated with the N_t and $N_p - N_t$, then the surprise is computed, finally the posterior becomes the prior and the cycle repeats until all the frames of the audio signal have been processed. In practice the equations (11) have to be modified by introducing a constant called "forgetting factor" ζ in order to avoid numerical overflows (Ψ and Γ functions grow with the order of $n!$).

$$\alpha_{k+1} = \zeta\alpha_k + N_t, \quad \beta_{k+1} = \zeta\beta_k + (N_p - N_t). \tag{13}$$

Normaly $\zeta \leq 1$, but it should be as close to 1 as possible, as it introduces a systematic error to the inference, while still preventing overflows. In our experiments $\zeta = 0.95$.
Originally the authors stated that the surprise should be computed as an integral over all possible models [8]. However we have observed empirically that the family of models with different priors but the same input data converges to a single, optimal set of parameters exponentially with the number of incoming data. Thus, to reduce the computational complexity we decided to use only one model per band with the initial uniform prior $(\alpha_0, \beta_0) = (1, 1)$

3.4. Selection of the onset candidates

The output of the model is a surprise matrix $S_M(frame, band)$ of real values $S(D, M)$. We are interested in finding the moments of the sudden increase of the surprise as we suspect them to be the most certain candidates for the onsets. However values of S_M are usually noisy, so for smoothing and amplifying of the desired growth of surprise, we correlate the surprise in each band separately with an appropriate spike.

$$spike = \{min(S_M), 2\bar{S_M}, \bar{S_M}\} \tag{14}$$

The spike signal is a discrete signal described as a set of only three consecutive values: $min(S_M)$ is the minimum value found in S_M and $\bar{S_M}$ is the arithmetic mean of all the elements of S_M. Then we sum the correlated matrix bandwise, obtaining a vector with a single value of summary surprise for each frame. Again we correlate it with a spike constructed as in Eq. (14), but taking the minimum and the mean values of the vector instead of the matrix. As a result we obtain a surprise curve over time. The local maxima of the curve, computed by finding the changes of its derivative are considered as the onset candidates.

3.5. Filtering of the onset candidates

Unfortunately, among the onset candidates there are still frames not containing onsets. It happens for two reasons, one is that the surprise points at the novel moments in the signal, the second reason is that the novel moment can be either a rapid increase but also a rapid decrease of a number of transient peaks. As a result, a surprise function contains two neighbouring local maxima, of which only one may be an onset. Another issue is that peaks associated with noise can also be surprising, and the model computes the surprise for each band separately. Thus, a mechanism that processes the information from events that occur simultaneously across the bands but at the same moment of time, must be introduced.

3.5.1. Filtering out unexpected drops of the number of transient peaks

In order to check if the number of transient peaks is increasing or decreasing, a local history of F_h frames preceeding the candidate, has to be checked. In our experiments we assumed $F_h = 4$, for hop size of 512 samples and window size of 4096 samples. For each band from the set of B bands, we construct a family of B Beta-Binomial models and each model has one of the possible B set of parameters, so the total number of models is equal to B^2. Each model from the family is updated with the data from F_h frames. The data are the number of the bands in each frame where the number of transient peaks has crossed the mean from F_h frames (within the same band). Among the models, the one with the smallest surprise is chosen as the one that supposely fits the data the best. From the model we compute the likelihood of the increase of the transient peaks for each of F_h frames, and if the frame with the maximum likelihood is the candidate frame, we keep the candidate, otherwise it is rejected.

3.5.2. Detection of the synchronous bursts

A Beta-Binomial model is used once more, here to estimate how many local maxima of the S_M are simultaneous. As in the above paragraph, the candidate frame must be processed with F_h history frames, because sometimes onsets are characterized by joint, synchronous bandwise, local maxima of the surprise matrix, but spanning across two consecutive frames. This seems to be an artifact introduced by the windowing, when the onset is placed in between the frames. Using more frames (here F_h) ensures that the onset can still be found even if the preprocessing-stage STFT window is relatively long (because of the well-known tradeoff between spectral and temporal resolution).

In this stage we use a simple but interesting property of surprise function that allows to estimate the "flatness" of the data. To illustrate the property with an example, let us imagine a person performing two series of throws of a coin (the possible outcomes: H-head, T-tail). The first series of let us say $N_{th} = 10$ throws looks as follows: H,T,H,T,H,T,H,T and, for the second series the outcome is H,H,H,H,H,T,T,T,T,T. In traditional Bayesian statistics, regardless of the prior beliefs we have before the person throws a coin, the posterior probabilities for both outcomes are close to 0.5, suggesting that the coin is fair. However, the posterior will not tell us anything about the changes of the distribution representing our beliefs about the fairness of the coin, along the consecutive N_{th} throws. But if we observe the curves of surprise, updated after each throw, we can see, that for the first series the function has a lot of peaks although the height of the peaks is dropping with time. For the second series, the surprise has only two peaks, one after the first throw and one after the fifth throw, but they are higher and have similar heights. The surprise drops with time with the speed proportional to N_{th}^{-1}.In our case the occurence of the local maximum of S_M in the band b and one of a two neighbouring frames $k, k+1, k \in [k_c - F_h, k_c]$, corresponds to throwing a head (for example) with k_c being the position of the onset candidate. We can measure the flatness as follows

$$F = \frac{\bar{S}(k)}{N_m} \tag{15}$$

where $\bar{S}(k)$ is the mean of the surprise over the frequency bands in the two frames and N_m is a number of local maxima of $S(k)$.We are interested in the highest F possible because the higher the F, the more synchronized are the peaks of S_M. Our experiments have shown that the condition $F > 3.5$ for acceptation of an onset candidate provides the best results. Local maxima of the S_M are shown in the Figure 2.

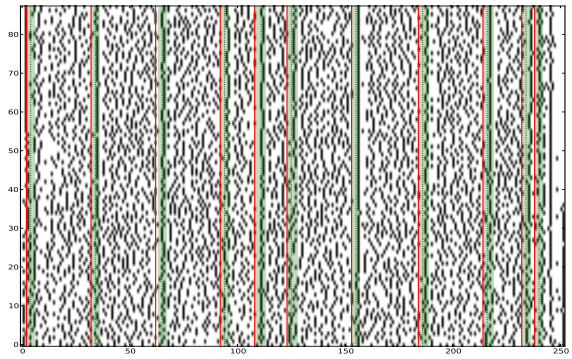

Figure 2: *The occurence of the local maxima of the surprise vs bands. The horizontal axis represents time in frames (512 samples / frame), whereas the vertical axis represents bands. The vertical bars are the ground truth onsets, surrounded by gray rectangles representing the tolerance window (50 ms). A local maximum is represented by a black spot. Between the onsets the distribution of the maxima is random, but at the onsets the maxima form vertical lines.*

4. Evaluation

4.1. Data set

The data set is a collection of 170 excerpts of music, each one has length of approximately 30 seconds. The collection comes from Digital Speech and Signal Processing research group at the Electronics and Information Systems department of Ghent University and it can be downloaded

at `http://dssp.elis.ugent.be` (see the "Downloads" section). The files are in the WAV format, mono and the sampling frequency is equal to 44100 Hz. It is annotated with the onset times. The set seems to be balanced in the sense of genre and rhythmic complexity.

4.2. Results

To estimate the performance of the model we use standard F-measure:

$$F = \frac{2PR}{P+R} \tag{16}$$

where

$$P = \frac{tp}{tp+fp} \qquad R = \frac{tp}{tp+fn}$$

where tp is the number of true positives, fp is the number of false positives and fn is the number of false negatives. As true positive we consider the time of a detected onset that differs from the time of the corresponding ground truth onset, less than 50 ms (a standard tolerance window used, for example, in the MIREX 2007). False positive is an onset that is detected but no ground truth onset lies closer than 50 ms to it. Finally, false negative onset is a ground truth onset for which no detected onset lies closer than 50 ms.

As we currently cannot compare our model with the original approach by Röbel [6], we have compared the performance of the model with an onset detector that is a combination of the High Frequency Content and Complex Domain methods, which has been being used for a long time in our lab [1]. To provide a reference to the current state of the art we also tested the detector made by Lacoste & Eck [5] as an example of the systems that apply machine learning mechanisms (in this case Artificial Neural Networks) for onset detection. The results are presented in the Table 1. The first two rows of the table illustrates the

Detector	Precision	Recall	F-measure
Lacoste & Eck, 2008	0.72	0.66	0.67
HFC+Complex domain	0.73	0.58	0.63
Bayesian, filtering enabled	0.68	0.52	0.57
Bayesian, filtering disabled	0.22	0.89	0.33

Table 1: *Performance of the onset detectors.*

performance of the two reference onset detectors. In the third row are the results obtained with the detector part only, up to the stage described in 3.4, in order to see how many onsets are missing in average. Finally, the last row of the table shows the performance achieved by the full system.

Figure 3 shows an example where the model output was perfect, reaching F-measure = 1.0. The input was a short passage of drums. As we can observe, the overall probability of an occurence of a transient peak is very low, except the in places where the onsets have been marked. Thus the onset events are novel and are marked correctly by the detector. Please also note the sudden growth of surprise just after the first onset. This is a moment of sudden drop of the number of transient peaks. As we see, it is also a surprising event, but the filter was able to distinguish it from the events that are real onsets.

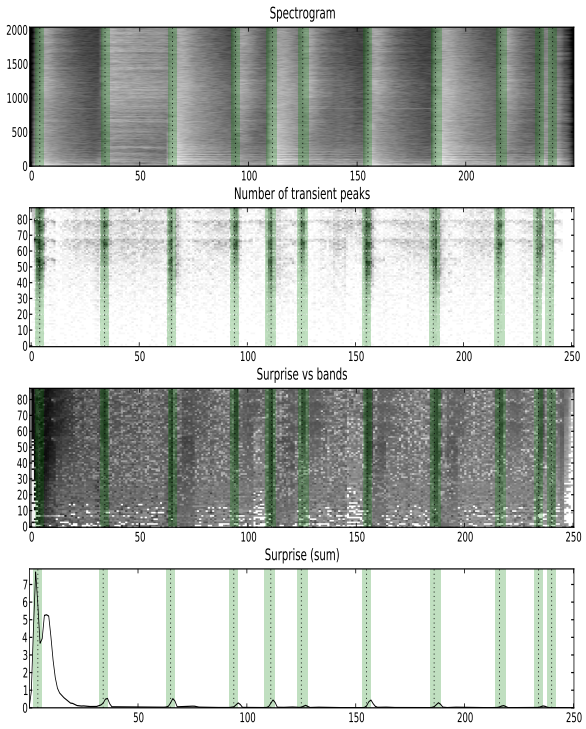

Figure 3: *Example output of the model. The horizontal axis represents time in frames (512 samples / frame). The vertical bars are the ground truth onsets, surrounded by gray rectangles representing the tolerance window (50 ms). The top image represents the spectrogram. The number of transient peaks per band is shown at the image below the top. The third image is the plot of the logarithm of the surprise (S_M matrix) vs. bands. The final, bottom plot is the surprise function. In the initial phase the surprise is always very high, regardless of the input data. However, the model learns with exponential speed, converging quickly to a stable state.*

5. Discussion and conclusions

Our aim was to prove that a marriage of Bayesian models with traditional Digital Signal Processing techniques can provide working solutions to some Sound and Music Computing problems. Our prototype solution, based on a very simple Bayesian model, came close with the performance to the standard solutions, entirely based on the DSP algorithms. The approach presented in this paper also introduces an unique look to the problem of onset detection, taking into account the timing information. At the bottom plot of Figure 3 (the surprise curve) we can observe that the height of the surprise function depends not only on the intensity of the event but also on time, so the onsets that are closer to each other show smaller changes of surprise. This property helped to distinguish between true onsets and false positives.

Although the overall performance leaves still room for improvement (see Figure 4), especially if compared to the current state of the art (see Table 1), we have shown that the onset detection done by the model can be perfect in case of drum passages, where the distributions of the transient peaks in case of onsets and inter-onset intervals differ significantly. Bayesian surprise allowed us to recognize synchronous events rapidly, without using complicated methods of pattern recognition like artificial neural networks or support vector machines. It is important to mention that the inference and the computation of surprise in

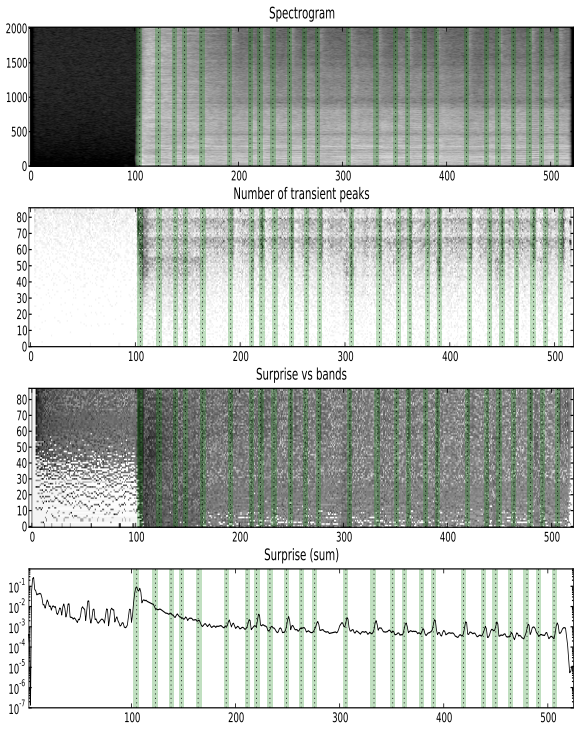

Figure 4: *Example output of the model for a rock music excerpt with a poor Signal-To-Noise ratio. The model performance is low, especially for the first bar (frames 100-200), as the distortions created by an electric guitar generate high amount of transient peaks. Only the beats are discriminant enough to be considered as surprising, mainly due the wide spectrum of some percussive instruments. For axes and plots description see Figure 3. The F-measure for the excerpt was equal to 0.62.*

the conjugate prior based models is very simple and robust. It does not require any conditional instructions in the code, except optional range checking of the input variables and each model is independent. These properties make the models particularly suitable for Graphics Processor Unit implemementations (using libraries like eg. OpenCL™or CUDA™).

In contrast to the Röbel's detector [6], our system has fewer parameters and it learns from the data in an entirely unsupervised way. That means it works seamless and produces repeatable results, not requiring tuning to a specific collection of data.

The system offers multiple, possible directions for improving the overall performance. Currently the distribution $Beta(\alpha_k, \beta_k)$ allows learning the most probable belief only, so the surprise depends mainly on the departure from the optimal α_k, β_k. That is why any peak of surprise can be a potential candidate for an onset. A natural solution of the problem are Hierarchical Bayesian Models [10]. Treating the parameters of the belief distribution α_k, β_k as random variables opens a possibility of learning deeper time dependencies at the price of significant complication of the inference algorithm. Moreover, the derivation of an exact, analytical formula for the surprise is usually impossible in this case. Another possibility is treating surprise as a random variable and classifying the surprising points depending on the distribution of this variable. However it seems to be difficult to classify due to its time dependency and to a wide range of possible values, therefore a machine learning method that could deal with these difficulties would have to be applied.

In any case, the most reasonable direction of the development seems to be incorporation of prior knowledge about music, reflected for example by the transition probabilities of the lenghts of the inter-onset intervals. Desain and Honing suggested that our rhythm perception may be driven by Bayesian learning [11] and we hope to obtain a major improvement by modeling higher level time dependencies present in music. The design of our model is particularly suited for such an extension.

6. Acknowledgements

This research is partially funded by the DRIMS project of the Spanish Ministry of Science and Education. We would also like to thank to our colleagues: Ricard Marxer for additional code, Ferdinand Fuhrmann and Ines Salselas for the reviews and the valuable remarks.

7. References

[1] Brossier, P., "Automatic Annotation of Musical Audio for Interactive Applications.", PhD Thesis, Centre for Digital Music, Queen Mary University of London, 2006.

[2] Quatieri, T., "Discrete-Time Speech Signal Processing. Principles and Practice.", ISBN 0-13-242942, Prentice Hall, 2002

[3] Rabiner, L.R.R. , "A tutorial on Hidden Markov Models and selected applications in speech recognition., Proceedings of the IEEE, vol. 77, pp. 257286, 1989.

[4] Abdallah, S., Plumbey, M., "Unsupervised onset detection: A probabilistic approach using ICA and a hidden Markov classifier.", In Cambridge Music Processing Colloquium, Cambridge, UK, 2003.

[5] Lacoste, A., Eck, D., "A supervised classification algorithm for note onset detection.", EURASIP Journal on Applied Signal Processing, 2007.

[6] Röbel, A., Onset detection in polyphonic signals by means of transient peak classification.,MIREX audio onset detection contest, 2007.

[7] Klapuri, A., "Multiple fundamental frequency estimation by summing harmonic amplitudes.", In Proc. ISMIR, pages 216–221, Victoria, Canada, 2006.

[8] Itti, L., Baldi, P., "Bayesian Surprise Attracts Human Attention.", In: Advances in Neural Information Processing Systems, Vol. 19 pp. 547–554, Cambridge,MA, MIT Press, 2006.

[9] Kullback, S., "Letter to the Editor: The Kullback–Leibler distance.", The American Statistician 41 (4): 340-341. JSTOR 2684769, 1987.

[10] Gelman, A., et al., "Bayesian Data Analysis, Second Edition", Chapter 5, Chapman & Hall/CRC., 2004.

[11] Desain, P. and Honing, H., The formation of rhythmic categories and metric priming., Perception-London, vol. 32, pp. 341–366, 2003.

[12] McEliece, R.J., Blaum, M., Farrell, P.G. and Tilborg, H.C.A.V., "Information, coding, and mathematics", Chapter 1, Springer, 2002.